looksmart DIRECTORY WEB ARTICLES Home ·
SEARCH all magazines FOR Search · Advanced Search · Help

YOU ARE HERE: Articles > Computer Technology Review > June, 2000 > Article

Print article Tell a friend Find subscription deals

InfiniBand: A Paradigm Shift From PCI.

Computer Technology Review, June, 2000, by Tom Heil

The transition from bus to fabric promises to be a major industry undertaking.

This article is the third in a three-part series. The second part appeared in the May issue of CTR.

InfiniBand is a compelling vision with the potential to revolutionize server architecture. As discussed last time, InfiniBand is driven by the pressing need to improve the scalability and availability of servers under tremendous pressure from Internet growth and it is backed by the industry heavyweights. The bus is history, right?

Not so fast. Despite its limitations (particularly in scalable environments) PCI is fast (getting faster with PCIX), simple, and cheap. InfiniBand is not just a significantly more complex technology, but a major computer architecture paradigm shift. The transition from bus to fabric promises to be a major industry undertaking. This article will explore some of the challenges InfiniBand faces and the impact these may have on how, where, and when the technology is adopted.

InfiniBand Adoption Challenges

The following is not all-inclusive, but simply meant to convey some of the complexities inherent in the move from today's bus architecture to tomorrow's switched fabric architecture.

* Specification Breadth and Complexity: Before converging on InfiniBand, the two competing switched fabric proposals--NGIO and Future I/O--addressed opposite ends of the price/performance continuum. NGIO was lean and cost-focused with an eye towards enabling relatively rapid PCI replacement in volume segments. Future I/O was performance-oriented and feature-rich, focused more on enterprise clustering than on immediate PCI replacement. It's no accident the Future I/O camp was the primary driver of PCI-X and the NGIO camp its primary detractor. This wide span of objectives is now under the common InfiniBand umbrella.

Is the vision too broad? Will features needed for higher end, lower volume applications put significant cost burden on lower end implementations? One of the key reasons behind PCI's success was its near exclusive focus initially on the high volume PC desktop. PCI shied away from features that might have made it more server and workstation friendly, but would have compromised PC cost structure. Now, practically all IA and RISC/Unix servers and workstations use PCI, not because it is optimal in

Sponsor
Ads by Google

VME Packa Solutions
by CG Mupa
Designed to
www.cgmupac.c

Clustering L servers
Cluster your servers with hardware - V
www.evidian.cor

Linux Clust Guide
From HP, for Techs: White case studies
www.hp.connect

PCI Express Architecture
Interconnect for embedde communicati applications
developer.intel.c

Content provide wi

THOMSON GALE

these segments, but because the economics of a technology backed by desktop volumes is hard to resist.

The IA market is tremendously interdependent. A complete InfiniBand solution will consist of piece parts--both hardware and software--from many, often competitive, vendors. Add this to the technology's intrinsic complexity and the risk of months, perhaps years, of specification "interpretation" issues, interoperability struggles, and the like is apparent.

* Viability of Unified SAN Model: InfiniBand is a unified System Area Network that promises to host any/all server I/O traffic types: LAN/WAN, cluster IPC, and storage. Although intuitively appealing and essential to PCI replacement, viability is not yet proven. There have been several "grand unification" attempts in the past--Asynchronous Transfer Mode (ATM) and Fibre Channel, for example--that, like InfiniBand, started out as "universal transports" destined to consolidate LAN, WAN, storage, voice, video, etc. and, in so doing, render all other wires obsolete. ATM and Fibre Channel found their niches, but neither came close to initial "all traffic on one wire" expectations. The devil is in the details. A universal transport cannot optimize every application. For example, there's inherent tension between large-block sequential I/O (typical of storage) and small-message random I/O (typical of IPC). Universal implies trade-offs, the impact of which won't be known for some time. The question is how much of a hit (relative to separate single-function networks) is the industry willing to pay for the benefits of a unified network?

1 · 2 · 3 · 4 · 5 | Next »

DIRECTORY WEB ARTICLES
SEARCH all magazines FOR Search · Advanced Search · Help
©2003 LookSmart, Ltd. All rights reserved. - About Us · Advertise with Us · Advertiser Log-in · Privacy Policy · Terms of Service

Return to article page
This story was printed from LookSmart's FindArticles where you can search and read 3.5 million articles from over 700 publications.
http://www.findarticles.com

InfiniBand: A Paradigm Shift From PCI.

Computer Technology Review, June, 2000, by Tom Heil

The transition from bus to fabric promises to be a major industry undertaking.

This article is the third in a three-part series. The second part appeared in the May issue of CTR.

InfiniBand is a compelling vision with the potential to revolutionize server architecture. As discussed last time, InfiniBand is driven by the pressing, need to improve the scalability and availability of servers under tremendous pressure from Internet growth and it is backed by the industry heavyweights. The bus is history, right?

Not so fast. Despite its limitations (particularly in scalable environments) PCI is fast (getting faster with PCIX), simple, and cheap. InfiniBand is not just a significantly more complex technology, but a major computer architecture paradigm shift. The transition from bus to fabric promises to be a major industry undertaking. This article will explore some of the challenges InfiniBand faces and the impact these may have on how, where, and when the technology is adopted.

InfiniBand Adoption Challenges

The following is not all-inclusive, but simply meant to convey some of the complexities inherent in the move from today's bus architecture to tomorrow's switched fabric architecture.

* Specification Breadth and Complexity: Before converging on InfiniBand, the two competing switched fabric proposals--NGIO and Future I/O--addressed opposite ends of the price/performance continuum. NGIO was lean and cost-focused with an eye towards enabling relatively rapid PCI replacement in volume segments. Future I/O was performance-oriented and feature-rich, focused more on enterprise clustering than on immediate PCI replacement. It's no accident the Future I/O camp was the primary driver of PCI-X and the NGIO camp its primary detractor. This wide span of objectives is now under the common InfiniBand umbrella.

Is the vision too broad? Will features needed for higher end, lower volume applications put significant cost burden on lower end implementations? One of the key reasons behind PCI's success was its near exclusive focus initially on the high volume PC desktop. PCI shied away from features that might have made it more server and workstation friendly, but would have compromised PC cost structure. Now, practically all IA and RISC/Unix servers and workstations use PCI, not because it is optimal in these segments, but because the economics of a technology backed by desktop volumes is hard to resist.

The IA market is tremendously interdependent. A complete InfiniBand solution will consist of piece parts--both hardware and software--from many, often competitive, vendors. Add this to the technology's intrinsic complexity and the risk of months, perhaps years, of specification "interpretation" issues, interoperability struggles, and the like is apparent.

* Viability of Unified SAN Model: InfiniBand is a unified System Area Network that promises to host any/all server I/O traffic types: LAN/WAN, cluster IPC, and storage. Although intuitively appealing and

essential to PCI replacement, viability is not yet proven. There have been several "grand unification" attempts in the past--Asynchronous Transfer Mode (ATM) and Fibre Channel, for example--that, like InfiniBand, started out as "universal transports" destined to consolidate LAN, WAN, storage, voice, video, etc. and, in so doing, render all other wires obsolete. ATM and Fibre Channel found their niches, but neither came close to initial "all traffic on one wire" expectations. The devil is in the details. A universal transport cannot optimize every application. For example, there's inherent tension between large-block sequential I/O (typical of storage) and small-message random I/O (typical of IPC). Universal implies trade-offs, the impact of which won't be known for some time. The question is how much of a hit (relative to separate single-function networks) is the industry willing to pay for the benefits of a unified network?

* Cost Structure: As previously mentioned, PCI's popularity in IA and RISC/Unix servers is based on the economies-of-scale associated with its dominance in desktops. There are segments where InfiniBand's value proposition will more than compensate for a higher technology cost structure; the ISP who's adding servers by the truckload, for example. But what about desktops, workstations, and entry servers? It's not clear what compelling value InfiniBand brings to these segments. Adoption then is going to be highly dependent on driving the cost delta between InfiniBand and PCI to near zero. (It should be noted too that bus performance may be hard to beat in this class system where latency is often more important than bandwidth scalability.)

The problem is that these entry systems represent, by far, the lion's share of system unit shipments. The High-Density Rack Mount (HDRM) environment where InfiniBand delivers clear value is growing fast, but is still small by comparison. InfiniBand may gain a foothold in HDRM, but if it does not branch out and dislodge PCI in high-volume segments, PCI will continue to outship InfiniBand by perhaps an order-of-magnitude or more. The market could fragment the way it did with SCSI and IDE disk drives. SCSI was always the more capable, scalable, feature-rich interface, but IDE drives still outship SCSI by an order-of-magnitude. Relatively speaking, only a small portion of the market needs and is, therefore, willing to pay a premium for SCSI. Likewise, InfiniBand is the more capable, scalable, feature-rich interface, but what is the premium and what segments will pay it? As long as PCI remains a high-volume opportunity, even scalable server OEMs may find it difficult to ship systems without at least a few PCI slo ts, to tap into the plethora of cheap PCI I/O.

So, can InfiniBand ever be as cheap as PCI? Maybe, but it will be difficult and will take time. Beyond the inherent costs of developing and deploying a new technology, the complexity of InfiniBand relative to PCI carries intrinsic cost. Granted, buses are pin-intensive, but logically, they are simple: a bunch of latches, buffers, and decoders managed by hardware state machines. In contrast InfiniBand is a full network protocol, requiring significant processing power and memory to manage queues, process messages, handle exceptions, etc. Over time, parts of the protocol stack will become automated to drive performance up and cost down, but initially, a lot of the work will be processor-based.

This may not be an issue on the host side since the host CPU can pick up the load. Targets, though, are another matter. Today's PCI-SCSI adapter is simple and cheap (often sub $100), little more than a single IC and connectors. In contrast, an early-market InfiniBand-SCSI TCA will likely consist of an InfiniBand-PCI front-end chip, a processor complex like i960 or PowerPC (with memory), and a PCI-SCSI back-end chip. This is fine for a RAID controller or storage router type product where the processor adds significant value, but cannot compete with the cost and performance of a simple PCI adapter. (A discrete processor-based TCA has to double buffer data, adding latency.) Ultimately, TCAs will have to be driven to a single chip before performance and electronics cost can approach PCI parity. Then, single chip versions of all required TCAs (SCSI, Fibre Channel, Ethernet, etc.) from multiple sources will be needed before hosts can safely do away with PCI slots.

Even if electronics achieve cost parity, a TCA board is inherently more expensive than a PCI adapter due to the TCA's mechanical canister that makes it customer replaceable. This is a terrific architecture attribute and an essential element of the InfiniBand value proposition, but does add to material and manufacturing cost.

Then, there's the switch. For InfiniBand to replace PCI, one would expect InfiniBand's "cost per port" to meet or beat PCI's "cost per slot." Today, switches connect boxes. For InfiniBand to replace PCI, switches must push beyond the box-to-box fabric and into the boxes themselves. A new class of inexpensive "edge" or "backplane" switch ICs is needed to bring the switched topology right to individual I/O devices. This is a new cost and architecture paradigm for switches. Recall that Fibre Channel Arbitrated Loop came about in response to the realization that switch cost was going to impede Fibre Channel adoption in more cost sensitive applications.

The most difficult cost structure issue, though, is economies-of-scale. If InfiniBand drove higher volume from day one, technology cost issues would take care of themselves, but it's the other way around. PCI is cheaper and will enjoy a significant volume edge for many years after InfiniBand's debut. PCI's unprecedented success gives it an economic inertia that will be difficult to overcome.

* Software: InfiniBand is as profound a paradigm shift to software as it is to hardware. Modular, "rack and stack" computing—where perhaps even, end-users can safely add or replace servers and I/O adapters—has a hardware and software component. Hardware is arguably the simpler problem. The enabling software infrastructure is a major undertaking, once again requiring cooperation between a broad base of participants. Imagine an HDRM environment with dozens of clustered servers, RAID controllers, LAN routers, and TCA enclosures. Someone replaces an InfiniBand-SCSI TCA. Who detects the event? Who is informed of the event? Which servers had access to the device? What if any of them had work in progress? How do the servers re-establish connection to the new device? If the new device is a newer version or from a different vendor, how do the servers get the appropriate driver? If multiple servers share the device, how do they coordinate access? These types of problems are not new, but they are complex and will requ ire significant cross-industry coordination (e.g., standards) to deliver the value, especially in the multi-vendor environment needed to drive down technology cost.

Take the transition from PCI to TCA device drivers, for example. Today, I write a driver to my PCI chip or adapter. Tomorrow, I will have to write it to some form of transport services layer that insures my message gets to my device on the other end of the network. Also I have to deal with things like connection management, lost packets, and device sharing, none of which I had to deal with in the simple, cozy world of close-proximity, in-box PCI adapters. What's more, my device on the other end probably sports a hefty firmware component given the shared, intelligent I/O nature of InfiniBand. The days of simple register reads and writes are over.

* Incumbent Technology Evolution: The toughest of all hurdles may be the unwillingness of incumbent technologies threatened by InfiniBand (and companies vested in them) to roll over and play dead. The burden-of-proof always rests on the challenger to demonstrate a value sufficiently compelling to overcome intrinsic incumbent advantages like maturity, multi-vendor availability, installed base, and supporting infrastructure and the bar never sits still. InfiniBand clearly sets a new watermark for architectural value, but in all practicality, it's still two to three years away. Meanwhile, PCI, Fibre Channel, SCSI, Ethernet, etc. will all press on per their respective roadmaps, steadily adding performance and value, and closing in on this new watermark. When NGIO first came on the scene, 66MHz PCI was the bandwidth watermark. Since then, this has doubled with PCI-X and could well double again before InfinBand achieves critical mass. Also, capabilities, like hot-plug and rack-friendly form-factors, are addressing many of the issues that make InfiniBand and so attractive. These

incremental solutions may not be as elegant as InfiniBand, but are available sooner at lower cost and use technology that the industry already knows. The longer it takes InfiniBand to reach critical mass, the more time incumbents have to close the gap. If, for example, Fibre Channel achieves 5Gbps or 10Gbps before InfiniBand has critical mass, will 2.5Gbps InfiniBand still be competitive?

InfiniBand Adoption

Given all this, how's InfiniBand likely to emerge? It probably helps to break the market into two categories: Internet-driven HDRM and traditional (everything else). By far, the most aggressive driver of HDRM is the Internet. More than any other segment, ISPs are crashing against the boundaries of today's architecture. The bus is especially problematic in this environment. Also, you don't need PCI slots in every server to take advantage of low-cost PCI adapters. It should be easy to provide dedicated "PCI adapter pool" servers in the room. If a slotless server needs a PCI adapter, it gets assigned one from the pool and "redirector" software passes I/O requests over InfiniBand to the pool server where the adapter and driver reside. This isn't pretty, but may suffice until all needed I/O is available natively and cost-effectively on InfiniBand.

Outside of Internet-driven HDRM (without which InfiniBand might not ever get off the ground), expect InfiniBand adoption to be much more gradual. Traditional clustered enterprise servers have to date been without a dominant standard IPC solution, meaning each OEM has had to carry the cost of its own proprietary solution. This seems like relatively low-hanging fruit for an early InfiniBand foothold. Over time, you would expect performance-critical RAID storage and LAN/WAN connections to move onto InfiniBand, especially as host chipset implementations prove themselves faster than PCI-based InfiniBand adapters. Until then, what's the point, since these systems will likely have PCI slots for many years to come. As long as "hybrid" systems--systems with native InfiniBand channels and PCI slots--are the norm, there will have to be a compelling value to put a particular function on InfiniBand, since PCI versions of that function will be more mature, more broadly available, and cheaper.

As you move down the traditional curve into stand-alone mid-range and finally entry servers and workstations, the InfiniBand value proposition gets progressively less compelling. Accordingly, adoption rates become increasingly dependent on maturity, solution availability, and cost structure parity. It's difficult to predict just how far down InfinBand will ultimately go. Will it go all the way down and truly obsolete the bus? Or will it stop short of entry server and workstation segments, resulting in a perpetual split where switched fabrics are the way to do scalable servers and buses the way to do entry servers and workstations? If InfiniBand doesn't capture the entry level, perhaps in time, some new I/O paradigm from the desktop or consumer world will.

The shift from PCI to InfiniBand is a complex undertaking that could take the better part of a decade to complete. A lot of things have to come together and it will be years before we know whether PCI can truly be retired or whether it will live on indefinitely, at least in some segments.

It's probably inevitable that at some point in the future, all I/O will exist at the end of a point-to-point link. The modularity, scalability, and ease-of-use attributes of switched fabrics and point-to-point topologies are just too attractive and the cost traditionally associated with such architectures comes down with each successive semiconductor technology generation. Given all the InfiniBand hype these days, it might be easy to overestimate its immediate market potential. It would be unwise, though, to underestimate the long-term potential of switched fabric architectures, like Infini-Band, to become pervasive throughout all of computing.

Tom Heil is the senior systems architect of the storage components division at LSI Logic (Fort Collins,

CO).

COPYRIGHT 2000 West World Productions, Inc. in association with The Gale Group and LookSmart.
COPYRIGHT 2000 Gale Group